A Statistical Study of the WPT-03 Corpus

Bruno Martins
Mário J. Silva

DI–FCUL TR–04–4

May 2004

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

Technical reports are available at `http://www.di.fc.ul.pt/tech-reports`. The files are stored in PDF, with the report number as filename. Alternatively, reports are available by post from the above address.

A Statistical Study of the WPT-03 Corpus

Bruno Martins
Mário J. Silva
Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal

bmartins@xldb.di.fc.ul.pt, mjs@di.fc.ul.pt

May 2004

Abstract

This report presents a statistical study of WPT-03, a text corpus built from the pages of the "Portuguese Web" collected in the repository of the tumba! search engine. We give a statistical analysis of the textual contents available in the Portuguese Web, including size distributions, the language of the pages, and the terms they contain.

1 Introduction

This study provides a statistical analysis of the textual contents on the Web page repository of the tumba! search engine [15]. More specifically, the source of information is the text extracted from a collection of documents from the "Portuguese Web", during the first semester of 2003. This roughly comprises all the pages hosted under the .PT top level domain (TLD), and other pages written in Portuguese and hosted in other TLDs (excluding .BR because most of these pages are also written in the Portuguese language).

The information presented in this study is of interest for the characterization of the textual contents of the Portuguese Web, as well as for future work within the scope of project tumba!. It is complemented by another report which provides statistics on the structure of the Portuguese Web [6].

The textual corpus is named WPT-03 and it is distributed by Linguateca (a resource center for the the processing of the Portuguese language – `http://www.linguateca.pt`) to researchers in the area of Natural Language Processing (NLP). For more information about the availability WPT-03, see the corresponding Web page at `http://xldb.fc.ul.pt/linguateca/WPT_03.html`.

The rest of this report is organized as follows: The next Section describes the WPT-03 corpus. In Section 3, we give statistics of the data in the corpus of web documents. Finally, Section 4 presents some conclusions.

2 Contents of the WPT-03 Corpus

The source of information for our study is a corpus of Web pages retrieved by the crawler of the tumba! search engine [5]. This snapshot of the Portuguese Web includes, for the most part, documents of types HTML and PDF, hosted in the .PT domain or written in Portuguese and hosted in the .COM, .NET, .ORG, or .TV domains.

The data was harvested and processed using the components from the XMLBASE Web database software, which includes the crawler, a Web content analyzer and a repository – see the project Web page at `http://xldb.di.fc.ul.pt/index.php?page=XMLBase`.

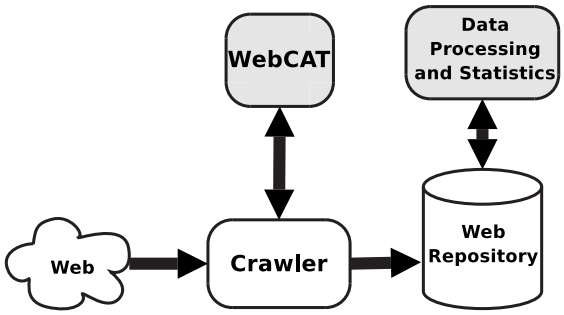

Figure 1: Overview of the XMLBase Framework

WebCAT is the tool responsible for parsing and analyzing the Web contents [8]. Among other things, it performs document format conversion, text extraction, and meta-data extraction. Both the original documents and the corresponding "textual versions" are maintained in Versus, a data repository for Web information [3]. All statistics are based on the corpus formed by the text documents stored in Versus.

The repository also contains meta-information about the documents, including for example the size, storage date, and language properties. Since there is no way of knowing the language in which the documents extracted from the Web were written, an automatic tool to perform this task had to be developed. This language "guessing" component is based on a well-known *n*-gram analysis algorithm [4], together with heuristics for handling Dublin Core meta-data (which may or not be available in the documents). In a controlled study, the algorithm presented a precision of about 91% in discriminating among 11 different languages [7].

A problem we faced concerns files in the PDF format – although most of the documents can be converted into plain text, the conversion tool sometimes fails in extracting the text, producing garbage as output instead of terminating with an error. Filtering this situations can be very hard. We currently exclude most of these faulty documents using a simple filter, which looks at the first characters of the file. However, this is not a perfect solution and many "garbage" documents are still included in the corpus.

Many of the presented statistics count "terms". We adopted a definition of "term" similar to that given by the Berkeley elib project – see the corresponding Web page at `http://elib.cs.berkeley.edu/docfreq`. According to it, terms are the sequences of the characters:

- a-z, A-Z, 0-9

- ASCII 150-160, 170, 181, 186, 192-214, 215-246, and 248-255 (Ű, Ů, Ÿ, Ź, Ž, IJ, İ, đ, §, ă, ł, ţ, ž, À, Á, Â, Ã, Ä, Å, Æ, Ç, È, É, Ê, Ë, Ì, Í, Î, Ï, Đ, Ñ, Ò, Ó, Ô, Õ, Ö, Œ, Ø, Ù, Ú, Û, Ü, Ý, Þ, SS, à, á, â, ã, ä, å, æ, ç, è, é, ê, ë, ì, í, î, ï, ð, ñ, ò, ó, ô, õ, ö, ø, ù, ú, û, ü, ý, þ, ß).

All other characters are regarded as term breaks. We differ from this definition in the way we handle hyphens. It is considered as a valid character of a term, when the next character is one of a-z or A-Z, in order to account that hyphens are essential characters in Portuguese, whereas in English they are mere punctuation marks (note that we still consider them as punctuation marks if they are not immediately followed by an alphabetic character). The definition of "term" adopted in this study includes therefore all sequences of the following characters:

- a-z, A-Z, 0-9;
- ASCII 45 (= "-");
- ASCII 150-160, 170, 181, 186, 192-214, 215-246 and 248-255.

3 Statistics of the WPT-03 Corpus

3.1 Document Statistics

The Portuguese Web snapshot analyzed in this study has 3775611 documents, collected between the 21st of March 2003 and the 26th of June 2003. Of these documents, about 68.6% (2590641 documents) are written in Portuguese.

Table 1 shows the average, median, and standard deviation of document sizes for WPT-03. Document size is measured in real size, text size and number of terms. Real size and text size are given in bytes, measuring the size of the document in the original format (HTML, PDF, ...) and converted into plain text, respectively.

	Real size	Text size	Number of terms
Average	24461	2886	438
Median	14672	1336	188
Standard deviation	54191	8240	1327

Table 1: Document size statistics

Figure 2 shows the distribution of document sizes measured in the number of terms. As in other corpora, the number of small documents is much higher and we conjecture that the distribution is identical. The distribution naturally follows Zipf's law [17, 11], as shown by the displayed trend-line.

3.2 Term statistics

3.2.1 Number and frequency of individual terms

Table 2 gives the total number of terms, the number of distinct terms, and the average and median number of occurrences of each distinct term. In order to abstract from differences in capitalization, all characters were converted into lower case before computing these statistics.

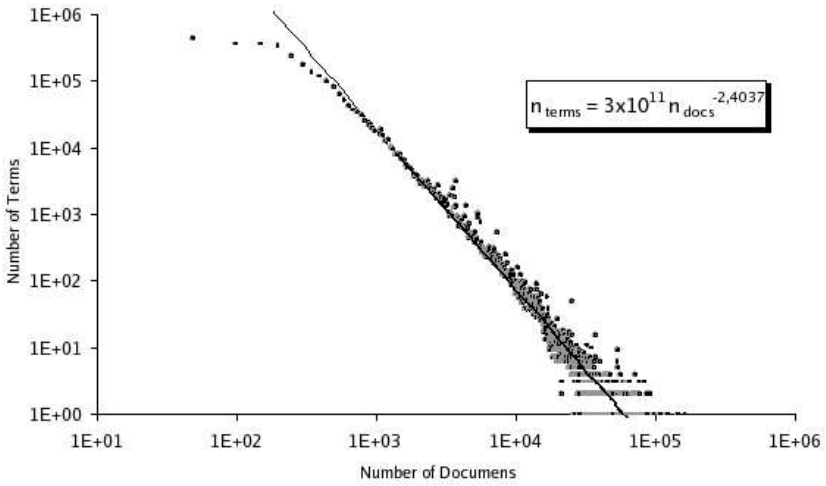

Figure 2: Document sizes in terms per document

	All Pages	Pages in Port. only
Total number of terms	1652645998	1208036873
Number of distinct terms	7880609	4066300
Average number of occurrences	210	297
Median number of occurrences	2	2
Standard deviation (# of occur.)	3428	42247
Average document frequency	865	128
median document frequency	1	1
standard deviation (doc. freq)	4496	5305

Table 2: Number of terms

The document frequency for terms, i.e., the number of documents in which a certain term appears (disregarding the number of occurrences in the document) is another important statistic. Since a substantial part of the documents are written in foreign languages, it is interesting to get some statistics for the terms occurring only in documents written in Portuguese. We therefore computed the frequencies considering both the full corpus and only the pages written in Portuguese.

Table 3 lists the 25 most frequent terms occurring in the corpus. Frequency is measured both in terms of the total number of term occurrences and document frequency, respectively. Most terms occurring in this list are candidate stop words in information retrieval systems for the Portuguese language.

3.2.2 Term size

We analyzed the average number of characters per term, regarding all terms occurring in the corpora and regarding all distinct terms. Additionally, we give the median and standard deviation. Once again the analysis is two-fold, with respect to all documents in the corpus and restricted to documents written in Portuguese. Results are given in Table 4.

All docs				Portuguese docs			
Term	Occ.	Doc.	Freq.	Term	Occ.	Doc.	Freq
de	58734369	de	2727182	de	55977484	de	2344461
a	35651699	a	2600458	a	29617180	e	2093982
e	27818162	-	2502955	e	26472070	a	2018658
-	22314054	e	2400583	o	21162843	o	1854189
o	21994175	do	2056158	do	16919378	do	1825455
do	17674236	o	2034963	-	15435398	-	1805399
da	15196359	da	1890699	da	14745024	da	1733306
que	14,659,562	os	1865796	que	1435160	para	1632962
the	14187020	para	1747633	em	10302921	em	1606919
1	12251543	em	1707146	para	9468453	os	1589019
em	10523210	com	1642535	os	8114119	com	1442380
para	9742012	no	1572418	com	7678022	que	1291286
os	8692680	as	1477125	1	7532463	por	1260588
com	8345476	1	1446037	um	6755990	um	1256344
0	8114739	que	1371220	no	6412220	no	1243250
2	8026747	2	1363302	por	5630947	na	1174801
no	7140563	por	1349638	não	5534784	as	1123699
um	6845245	um	1294310	as	5383637	dos	1072318
as	6800040	na	1240599	dos	5363523	uma	1053919
of	6427217	3	1183339	uma	5339622	não	1040530
and	6085323	s	1152840	2	5080441	ao	1040371
to	5934084	dos	1132724	na	5041565	todos	1037091
por	5825647	pt	1106868	é	4630006	1	1007115
não	5608615	uma	1102931	se	4351434	ou	950273
dos	5497639	todos	1090087	ou	4284627	2	944924

Table 3: Most Frequent terms

	All terms		Distinct terms	
	all docs	only port.	all docs	only port.
average	4.8840	4.9680	8.9780	8.7400
standard deviation	4.4762	3.7930	39.6400	20.6860
median	4	4	7	8

Table 4: Term size

Figure 3 shows the distribution of term size (regarding all terms of the corpora). Approximately 99% of the terms are shorter than 15 characters, and a major part of those longer than 15 characters are due to “garbage” in the corpus and the problem of extracting the text from PDF files mentioned above.

3.2.3 E-mail addresses, numbers and hyphen statistics

To improve the search engine’s handling of queries, it was interesting for us to analyze the frequency of things like e-mail addresses or numeric terms.

Numeric terms, as the name suggests, consist solely of numeric characters. As for e-mail addresses, they are of the form X@X.X, where X stands for a non-empty alphanumeric sequence plus the characters “-” and “_” (See the Internet RFC822 - Standard for the format of ARPA Internet text messages). Although each e-mail address counts as several terms in all other statistics (the separators are seen as punctuation), here they are seen as atomic units.

Finally, hyphenated words are terms where one character is a hyphen, as defined

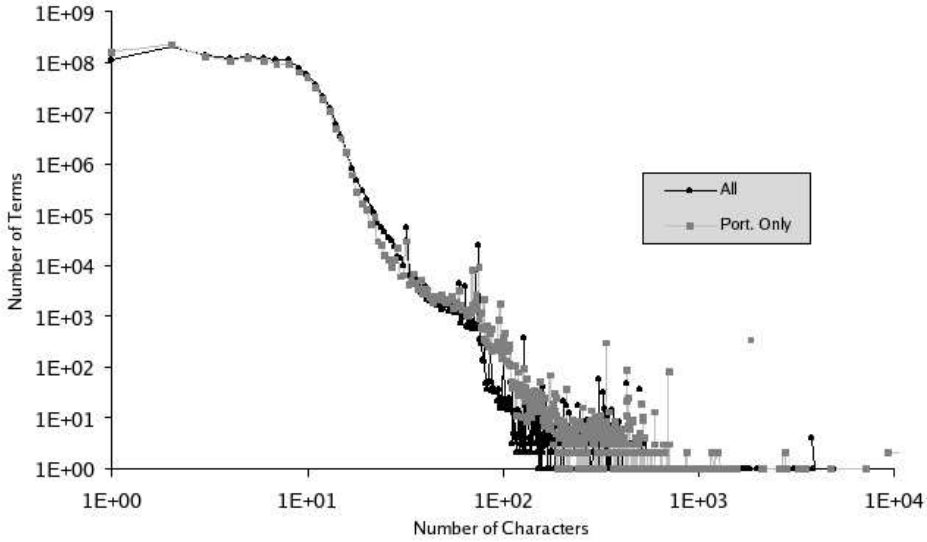

Figure 3: Term size in characters per term

in Section 3.1. Counting hyphenated terms is important as, depending on their frequency, it may be more interesting for the search engine to consider them as separated sequences of terms.

Table 5 shows the number of occurrences, the average number of occurrences for each distinct term in the collection, and the average size (in number of characters) of e-mail addresses, numeric terms and terms containing hyphens. The weighted average size also refers to the number of characters, but considering the number of all occurences, instead of only the distinct ones.

	E-mail addresses	Numeric terms	Terms with hyphens
Number of All Occ.	1264939	146136400	52499281
Number of Dif. Occ.	203638	570406	1510253
Average of occ.	6.21	256.20	34.76
Average term size	21.48	7.44	11.76
Average term size (weighted)	20.44	2.13	5.87

Table 5: Special terms

3.2.4 Morphology of the terms

In order to get an idea of the morphology of the terms occurring in the corpus, we used the jspell [16] morphologic analyzer. This allowed us to relate base forms of words to inflected variants, and find out which syntactic categories the terms in the corpus belong to.

After excluding all terms containing numeric characters, we obtained 1884932 distinct terms (regarding only the Portuguese documents). 429937 of these can be analyzed morphologically using jspell. To reduce ambiguity, we only accept a solution if

the lemma resulting from undoing inflection is contained either in the WPT-03 corpus itself, or in the CetemPublico corpus (see Section 3.3).

Of the 429937 terms that can be analyzed, 179778 (41.81%) are unambiguously analyzed as both nouns and adjectives, 137270 (31.93%) as verbs, 13932 (3.24%) as adjectives and 10322 (2.40%) as nouns. Furthermore, 71321 (16.59%) terms are ambiguous between verb and noun/adjective, 7117 (1.66%) between just noun and noun/adjective and 2342 (0.54%) between adjective and noun/adjective. 7855 (1.83 %) terms are ambiguous in other respects.

In the future, we plan on using other tools to enhance the morphology analysis in WPT-03, such as the PALAVROSO morphologic analyzer [9] or a good parts-of-speech tagger trained for the Portuguese language [1, 2].

3.3 Inter-corpora Statistics

This Section provides statistics comparing the tumba! corpus against CetemPublico [13, 12]. In the future, we plan to cross WPT-03 with other available corpora of Portuguese text, giving a more extended analysis.

3.3.1 CetemPublico

To measure the coverage of the dictionary used for spelling correction in tumba!, we analyze the appearance of terms in the corpus that are contained in the spelling dictionary. As the dictionary contains all the terms that appear in the CetemPublico corpus, this statistic not only provides information about correctly spelled terms, but also about the overlap of the CetemPublico and the tumba! corpora. Note that the correction of the terms can not be 100% assured, as the CetemPublico corpus used to build the spelling dictionary contains itself errors.

A substantial part of the terms differ only by the use of accents (i. e., replacing for example *á* by *a*). For that reason, in the statistics that compare these corpora, we provide on Tables 6 and 7 two result sets: one considering accented characters, and the other ignoring them. The meaning of each line on both tables is as follows:

#WPT-03 terms in CP (distinct) indicates the number of distinct terms in the WPT-03 corpus that also occur in the CetemPublico corpus; the percentage represents how many of the distinct terms of WPT-03 also appear inside the CetemPublico corpus.

 # CP terms in WPT-03 (distinct) indicates the number of distinct terms in CetemPublico that also occur inside the WPT-03 corpus; this is the same number as above, but the percentage is slightly different.

 # WPT-03 terms in CP (total) indicates the total number of terms in the WPT-03 corpus that also occur in CetemPublico; the percentage represents how many of all the terms in WPT-03 also occur inside the CetemPublico corpus.

 # CP terms in WPT-03 (total) indicates the total number of terms in the CetemPublico corpus that also occur in WPT-03; the percentage represents how many of all the terms in CetemPublico also occur inside the WPT-03 corpus.

Note that whereas almost all of the terms in CetemPublico also occur in the WPT-03 corpus, only 60% of the terms from WPT-03 appear in the CetemPublico corpus.

	All docs	Only port.
#WPT-03 terms in CP (distinct)	153729 (1.95%)	152641 (3.75%)
# CP terms in WPT-03 (distinct)	153729 (3.46%)	152641 (3.43%)
# WPT-03 terms in CP (total)	4213578 (94.77%)	4212486 (94.74%)
# CP terms in WPT-03 (total)	984033934 (59.54%)	897616340 (74.30%)

Table 6: Overlap with CetemPublico (counting all characters)

	All docs	Only Port.
#WPT-03 terms in CP (distinct)	150157 (1.91%)	148913 (3.66%)
# CP terms in WPT-03 (distinct)	150157 (38.89%)	148913 (38.57%)
# WPT-03 terms in CP (total)	4221291 (94.94%)	4219958 (94.91%)
# CP terms in WPT-03 (total)	1003575179 (60.73%)	907834378 (54.93%)

Table 7: Overlap with CetemPublico (ignoring accentuated characters)

This is, at least partly, due to the amount of documents written in languages other than Portuguese, and also to CetemPublico being much "cleaner", i.e., it contains less terms including numeric characters or "garbage" text. Previous studies have already indicated that while Web corpora have advantages in quantity (more "live" language information, more words and case-frames that newspaper corpus, ...), they are usually a lot noisier [14].

3.3.2 Postal Codes

Having an idea of the amount of geographic entities that are present in the WPT-03 corpus would be very interesting for us in the context of project tumba!. We used a list of Portuguese postal codes to find out which and how many "geographic" names appear in the text. The list is provided by CTT (Portuguese Post Office) and can be downloaded from `http://codigopostal.ctt.pt/pdcp-files/todos_cp.zip`. It contains not only postal codes, but also city, street and district names (277980 names of geographic entities overall).

	All Postal Codes	Distinct Postal Codes
WPT-03	683458	33799
CTT	236924	170549

Table 8: Postal Codes found in tumba! and in CTT list.

In the analysis, we considered all terms in the form XXXX-XXX as postal codes, with X being a numeric character. Tables 3.3.2 and 3.3.2 show the statistics for postal codes occurrences. Near 1/6 of all Portuguese Postal Codes appear in the WPT-03. We can speculate that these are the Postal Codes for areas where many business and commercial entities are located. The amount of CTT postal codes in the WPT-03 that also occur in the CTT database should be 100%, but 17-19% of the Postal Codes in WPT-03 are infact invalid.

3.3.3 Geographic Entities

To have an idea of the richness of WPT-03 on geographical references, we searched the corpus for such information. For this purpose, we did a case-insensitive search on the

	All Docs
CTT Postal codes in WPT-03 (distinct)	27695 (81,94%)
WPT-03 Postal Codes in CTT (distinct)	27695 (16,24%)
CTT Postal codes in WPT-03 (total)	567326 (83,01%)
WPT-03 Postal Codes in CTT (total)	60885 (25,70%)
Average of occ.	2,20

Table 9: Statistics of Postal Code occurrences.

308 Portuguese municipalities.

As many Portuguese geographic names consist of more than one word, we need to group individual terms, in order to provide statistics on the geographic entities identified in the corpus. To locate these entities, we use a simple algorithm that looks at all matches of those geographic names in the "word-grams" from WPT-03.

The total number of geographic entities identified in the corpus using this method is 8147120. The ten most frequent are given in Table 10, along with the overall number of occurrences.

Geographic Name	Number of Occurrences
lisboa	1034268
porto	651108
coimbra	307881
guarda	198436
aveiro	192804
braga	186410
almeida	142591
leiria	121280
faro	111028

Table 10: Most Frequent geographic names

This was only a crude approach to measure the amount of geographic references, and the results are not conclusive. For instance many Portuguese proper names (especially people's names) are also geographic names, and they were identified in this study as geographic references. In the future, we plan on conducting a much more accurate analysis of the occurrence of this information on WPT-03, using specific software for accurate named entity recognition.

4 Conclusions

We used the tumba! repository to construct a textual corpus from the pages of the "Portuguese Web", denominated WPT-03. The corpus was then analyzed using common statistical techniques from corpus linguistics [10].

This study was motivated by our interest in finding more about the textual contents of the tumba! repository, including both information about the documents (their size, language, ...) and the terms contained in the documents. With this data we can better model the capacity and the algorithms of the tumba! search engine, and in tandem provide insights to a large corpus in natural language that are of interest to other researchers.

Specially interesting is comparing WPT-03 with several other corpora made available through Linguateca. WPT-03 contains the more or less colloquial language found on Web pages, whereas the Linguateca corpora have mostly the more formal language found in newspaper articles.

It will also be interesting to repeat this study regularly and track the evolution of the data – most probably the most frequent terms (apart from function words, of course) will change over time. The study could also be carried out using different sub-corpora, in order to find differences and similarities for different Web "communities".

Finally, a complementary study on the logs for the queries submitted to tumba! would also be very useful, in order to understand the way Portuguese users search for information on the Internet, and if the information they are looking for is widely available or not.

5 Acknowledgments

We wish to thank Ruth Fuchss, for developing the initial scripts to compute statistics over the corpus, and Daniel Gomes for doing most of the work in harvesting WPT-03. Nuno Cardoso did several optimizations on the original scripts, and also helped on reviewing early drafts of this paper. Finally, a special thanks goes also to Diana Santos, for providing us with valuable insights and suggestions.

References

[1] T. Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP-00, the 6th Conference on Applied Natural Language Processing*, 2000.

[2] E. Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, the 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy, 1992.

[3] J. P. Campos. Versus: a web data repository with time support. DI/FCUL TR 03–08, Department of Informatics, University of Lisbon, May 2003. Masters thesis.

[4] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *Proceedings of SDAIR-94, the 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, Nevada, U.S.A, 1994.

[5] D. Gomes. Tarântula – sistema de recolha de documentos da Web. Technical report, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa, August 2001. Report of the traineeship done by the author at the LaSIGE `http://lasige.di.fc.ul.pt`. In portuguese.

[6] D. Gomes and M. J. Silva. A characterization of the Portuguese Web. In *Proceedings of the 3rd ECDL Workshop on Web Archives*, Trondheim, Norway, August 2003.

[7] B. Martins and M. Silva. Language identification in Web pages, 2004. (To appear).

[8] B. Martins and M. Silva. WebCAT: A Web content analysis tool for IR applications, 2004. (To appear).

[9] J. C. D. Medeiros. Processamento morfológico e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, 1995.

[10] M. P. Oakes. *Statistics For Corpus Linguistics*. Edinburgh University Press, February 1998.

[11] V. Poosala. Zipf's law. Technical Report 900 839 0750, Bell Laboratories, 1997.

[12] D. Santos and P. Rocha. Evaluating cetempúblico, a free resource for portuguese. In *Proceedings of ACL-2001, the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, July 2001.

[13] D. Santos and L. Sarmento. O projecto AC/DC: acesso a corpora / disponibilização de corpora. In A. Mendes and T. Freitas, editors, *Actas do XVIII Encontro da Associação Portuguesa de Linguística*, pages 705–717, October 2002.

[14] Y. Sekiguchi and K. Yamamoto. Web corpus construction with quality improvement. In *Proceedings of IJCNLP-04, the 1st International Joint Conference on Natural Language Processing*, pages 201–206, 2004.

[15] M. J. Silva. The case for a portuguese web search engine. In *Proceedings of ICWI-2003, the IADIS International Conference WWW/Internet 2003*, November 2003.

[16] A. M. Simões and J. J. Almeida. `jspell.pm` – um módulo de análise morfológica para uso em processamento de linguagem natural. In *Actas da Associação Portuguêsa de Linguística*, pages 485–495, 2001.

[17] G. K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Reading, Massachussets, U.S.A., 1949.